Recommending additional study materials: Binary ratings vis-à-vis five-star ratings

Juha Leino

University of Tampere / School of Information Sciences / TAUCHI
Kalevantie 4, 33014 Tampereen yliopisto, Finland
juha.leino@uta.fi

As various recommender approaches are increasingly considered in e-learning, the need for actual use cases to guide development efforts is growing. We report on our experiences of using non-algorithmic recommender features to recommend additional study materials on an undergraduate course in 2009–2011. The study data comes from student e-questionnaire replies and actual click-by-click use data. Our discussion centres on using binary (useful/not useful) rating scale (2009–2010) vis-à-vis five-star rating scale (2011). Using five-star scale to increase the complexity of the rating decision significantly reduced dishonesty (rating items without viewing them), but at the price of fewer ratings overall and increased complexity of interpreting the ratings. In addition to explaining how ratings and other factors inter-influenced item-selecting, we also discuss how different scales (binary and five-star) affect the rating behaviour in e-learning and how the five-star rating distributions in e-learning relate to those in other domains. Furthermore, we discuss two models, *high-quality approach* and *low-cost approach*, of employing non-algorithmic recommending features in e-learning that emerge from our findings. The findings provide the field with insight into the actual dynamics of using recommender features in e-learning. Moreover, they provide practitioners with actionable information on dishonesty.

Recommender systems, ratings, scales, binary, five-star, e-learning, honesty.

1. INTRODUCTION

Recommenders help us with information overload in two partially overlapping ways, by helping us find salient items (for example, books we might be interested in) and by supporting decision making (for example, which book to buy) (Swearingen and Sinha, 2001; Schafer et al., 1999). Recommender systems consist of one or more recommender features that span from comments/reviews (often shown as entered) through ratings (shown individually or as aggregates) to prediction-computing (Leino, 2011; Schafer et al., 1999).

While recommenders were first applied in such areas as e-commerce, today they are also starting to be increasingly employed in education and e-learning (Ghauth and Abdullah, 2010; Tang and McCalla, 2004). Although recommenders serve similar purposes in e-commerce and e-learning, differences between the domains make transferring recommenders directly from one to another challenging (Drachsler et al., 2009; Tang and McCalla, 2004). Even within e-learning, various recommending tasks and contexts place differing demands on recommenders (Ghauth and Abdullah, 2010; Drachsler et al., 2009). Currently, there is a need for case studies of using recommenders in e-

learning, as only very few systems have been evaluated with trials involving actual users in authentic use contexts (Manouselis et al., 2011).

This article discusses our experiences of using non-algorithmic recommender features on an undergraduate course on user-centred design (UCD) to recommend additional study materials to students in 2009–2011. A non-algorithmic approach was adopted because the use period (one semester) and low numbers of items, students and interactions did not allow using algorithmic approaches.

A large majority of college/university students already complement course materials with online sources and are willing to share these with fellow students (Hage and Aïmeur, 2008). Consequently, the author added a Lecture Slide and Reading Materials (LSRM) page to the course website (requires logging in) to allow students both to add salient reading materials to the page and to evaluate them as a community of peers with ratings, comments and tags, thus aggregating communal knowledge. The goal was both to help students find high-quality materials on UCD, i.e. *find good items* (Herlocker et al., 2004), and to encourage them to read more widely on UCD.

Our discussion is based on student e-questionnaire replies and actual click-by-click use data. The 2009 experiences of binary ratings and commenting have already been discussed in Leino (2011) and Leino (2012). Consequently, here we focus on experiences of five-star ratings (2011) vis-à-vis binary ratings (2009–2010). As tagging (2010–2011) was used little and failed to provide much value, we omit it from the discussion.

Overall, students felt that the recommender features increased interactivity and sociality, and added interest to the otherwise boring list of materials. By increasing the complexity of the rating decision, five-star rating scale reduced the number of dishonest ratings (ratings made without viewing the material) but simultaneously it also reduced the overall number of ratings. Interestingly, rating distributions differed from the J-shaped distributions prevalent in e-commerce (Hu et al., 2009). The results suggest two models for using non-algorithmic recommending features in e-learning, *low-cost approach* (increasing the number of evaluations by lowering the cost of contributing) and *high-quality approach* (increasing the quality of evaluations by raising the cost of contributing).

The results contribute to the field by showing how to decrease dishonesty in ratings in e-learning. Also, we are the first to discuss how different scales (binary and five-star) affect the rating behaviour in e-learning and how the five-star rating distributions in it relate to those in other domains. Understanding the differences in the use practices in different domains is important to facilitate transferring experiences between domains. Finally, our results provide currently lacking insight (Manouselis et al., 2011) into the actual use dynamics of recommender features in e-learning.

2. BACKGROUND

Having users explicitly rate items is considered a reliable way to collect user preferences (Ghauth and Abdullah, 2010). However, as ratings in e-commerce are typically overwhelmingly positive and their distributions bimodal and often J-shaped, their ability to reflect true item quality has been questioned (Hu et al., 2006; Hu et al., 2009; Talwar et al., 2007). In fact, contributions may be coming mainly from highly opinionated users (Talwar et al., 2007). Hu et al. (2006) suggest a *brag and moan model* to explain the prevalence of bimodal distributions while Hennig-Thurau et al. (2004) similarly suggest that strong consumption experiences may result in *expressing positive emotions* or *venting negative feelings*.

There are few guidelines to choosing between different rating scales (Sparling and Sen, 2011) despite rating scales affecting the distribution of ratings; some scales produce higher and some

lower ratings on the same item (Gena et al., 2011). Also, the interface implementation of the rating scale affects the equation (Cosley et al., 2003). Cena et al. (2011) consider various scales and their visual metaphors to have emotional connotation.

Contradicting Cosley et al. (2003), Gena et al. (2011) found that ratings on different rating scales do not scale linearly, thus preventing normalization between scales. This underlines the need to understand better the effects of different rating scales. Moreover, rating distributions appear domain-specific (Gena et al., 2011; Sparling and Sen, 2011).

Finally, showing predictions in the interface influences people's opinions on items and can consequently affect item-selecting and rating behaviour (Cosley et al., 2003). Talwar et al. (2007) also found that past ratings and reviews can influence the future ones by creating expectations to which users react based on their actual experiences.

3. STUDY SETTING AND DATA COLLECTING

Data for this study was collected in 2009–2011 on an undergraduate-level course on UCD (Table 1) that consisted of seven lectures (Sept–Oct) and fourteen practice sessions (Sept–early Dec). The course-work consisted of a design assignment (DA), ten smaller assignments and presenting the DA in class. All this work was done in small groups of mostly three students while online activity—adding and evaluating materials—was individual.

Table 1: Students

	Students in the course		Questionnaire respondents	
	No. of students	No. of females	No. of students	No. of females
2009	32	10 (31%)	23 (72%)	7 (30%)
2010	55	14 (25%)	22 (40%)	8 (36%)
2011	37	8 (22%)	19 (51%)	5 (26%)
Total	124	32 (26%)	64 (52%)	20 (31%)

Table 2: Course grading

	2009	2010–2011
Design assignment	80%	70%
Smaller assignments	20%	20%
Online activity (LSRM)	Max. -10%	10%
Attendance (extra)	10%	10%
Max. total	110%	110%

In 2009, students faced a punishment of max. 10% of the grade if they failed to do the required online activity (add one material and rate five). In 2010–2011, the stick was turned into a carrot by moving ten percentage points from the DA to online activity (Table 2). Simultaneously, the activity requirement was raised to adding two materials and rating five.

Commenting and tagging were always non-compulsory. Students were not examined on the additional study materials so reading them was in that sense voluntary.

Our discussion is based on e-questionnaire replies and actual use data. About half the students (52%) filled out the questionnaire (Table 1) that consisted of the following sections: Material quality, Adding materials, Ratings, Commenting, Tags (2010–2011), Tools in general and General impressions. In 2009, a five-point Likert-style scale was used for evaluative statements. In 2010–2011, a seven-point interval scale was adopted instead (1=Strongly negative; 7=Strongly positive). Most evaluations allowed textual commenting. Also, there were numerous open-ended questions. Movie tickets were raffled among the students who filled out the questionnaire.

The LSRM page recorded virtually every click into a database. This use-log data provides a rather complete picture of the student activity on the page and balances saying (questionnaire) with actual doing (log data).

Although studies of authentic use such as this are more subject to confounding factors than laboratory studies, they provide insight into the user experience and actual user behaviour. In fact, user

experience is the decisive factor as far as recommender systems are concerned (Herlocker et al., 2004; Konstan, & Riedl, 2012). Moreover, although our results come from a case study where the numbers of users and items is low, they also have implication to scenarios where the numbers are significantly bigger, as for example improving rating honesty is important in any e-learning scenario.

4. LSRM PAGE

The LSRM page was implemented with HTML, PHP, JavaScript and AJAX. Using AJAX allowed all interactions—adding materials and rating, commenting and tagging them—to take place without reloading the page. All activity was anonymous in all three years.

Material links were added to lectures (under which they were organized on the page). Clicking *Add a reading material* link opened a form with fields for material title and URL. The added link was placed on top of the materials for the lecture.

In 2009, the LSRM recommender system consisted of binary ratings and commenting (Figure 1). The binary rating approach—Yes and No buttons to respond to the question ‘Did you find this material

2009

1/1 students have found this material useful. Did you find this material useful?
Taking your Site from One to One Million Users by Kevin Rose [2009-10-17 14:16]
 This material has not yet been commented. [Add comments](#)

3/3 students have found this material useful. Did you find this material useful?
7 Rules for Mixing Multiple Fonts in Good Web Design [2009-10-11 19:53]
 This article has **1 comments**. [Show comments](#) | [Add comments](#)

2010

2/2 students have found this material useful. Did you find this material useful?
James Hobart: Principles of Good GUI Design
 Tags: [Add tags](#)
 This material has **1 comments**. [Show comments](#) | [Add comments](#)

11/11 students have found this material useful. Did you find this material useful?
Usability Methods
 Tags: [Add tags](#)
[user observations](#), [design](#)
 This material has not yet been commented. [Add comments](#)

2011

☆☆☆☆☆ (0 ratings) Rate this link: ☆☆☆☆☆
Real or Imaginary: The effectiveness of using personas in product design (by Frank Long)
 Tags: [Add tags](#)
[scenario](#), [persona](#)
 This article has **1 comments**. [Show comments](#) | [Add comments](#)

★★★★☆ (3 ratings) Rate this link: ☆☆☆☆☆
An introduction to personas and how to create them
 Tags: [Add tags](#)
[persona](#)
 This article has **2 comments**. [Show comments](#) | [Add comments](#)

Figure 1: LSRM user interface in 2009 (top), 2010 (middle) and 2011 (bottom) (translated from Finnish)

useful?’ to make the rated aspect explicit—was adopted to allow expressing both liking and not liking. Students were able to change their ratings but not to delete them.

Commenting was decoupled from rating to keep the rating cost as low as possible (one click). Clicking *Add comments* opened a form with fields for comment title and content. While students could not rate the materials they had added, they could comment them to enable discussing.

Tagging was added in 2010 but otherwise the interface stayed the same (except that the time the material was added was no longer displayed).

In 2011, the binary rating scale was replaced with a more granular one to allow evaluating items more exactly and to increase the complexity and therefore the cost of the rating decision (Sparling and Sen, 2011) to reduce the number of dishonest ratings (ratings made without viewing the material) and to increase both perceived and actual reliability (Cosley et al., 2003; Sparling and Sen, 2011). A five-star rating approach was adopted because users prefer it (Gena et al., 2011; Sparling and Sen, 2011). Furthermore, stars provide an easy and familiar way to visualize rating averages.

The interface separated the displaying of the average of all star ratings (top-left corner of a material block) from the rating interface (top-right). Rating was done by moving the mouse over the stars to ‘light up’ (turn yellow) stars and then clicking. Ratings were displayed at the precision level of half stars and made at the level of full stars.

5. STUDENT VIEWS ON AND USE OF THE RECOMMENDING FEATURES

Overall, students evaluated the recommending features positively (Table 3). Features were also commented on positively—‘*They brought interactivity. Without such stimuli I’d have visited the link page less*’ (2010)—and were seen as ‘pleasant’ (2010) and as ‘*absolutely better than a mere list*’ (2010) that would have been ‘*immensely boring*’ (2011). While some students felt that the tools did not help them much, they were not against having them: ‘*Although I didn’t much use the materials or the tools, I still believe that they were helpful to many others*’ (2011).

The features were seen as encouraging the reading of materials and increasing sociality: ‘*...it gave me a feeling that somebody actually reads materials for real and reading the viewpoints of others is important for me*’ (2011). Perhaps the best indication of student appreciation was the fact that several mentioned wishing to have similar tools available also in other courses.

Table 3: Student evaluations of the recommending features (standard deviation in parentheses) (2009 numbers adjusted from 5-point scale to 7-point scale)

Feature evaluated	2009	2010	2011
Tagging	-	4.82 (.96)	4.13 (1.09)
Rating	5.29	5.68 (1.25)	5.11 (1.03)
Commenting	5.18	5.55 (1.18)	5.79 (1.03)
Adding materials	5.29	5.36 (1.18)	5.11 (.99)
Overall ease-of-use	5.96	5.95 (1.09)	5.06 (1.16)

Overall, recommender feature use was largely determined by compulsoriness (Table 4). About 30% of the students made few contributions, most did more or less what was required and a few did more, some in fact significantly more. The 2009 and 2010 numbers are very similar (see also Tables 5 and 6 below), underlining that the same interface produced similar results despite the stick (punishment of max. 10% of the grade) having been turned into a carrot (online activity forming 10% of the grade). This highlights the importance of the interface to the user behaviour and suggests that the differences between 2009–2010 and 2011 are largely because of the changes in the interface/rating scale and not random variation due to different groups of students.

Table 4: Use of the rating and adding material features
(¹In 2009, students were required to add one material while in 2010–2011, two; ²One rating by teaching assistant; in 2010–2011, all ratings were by students.)

		2009	2010	2011
Reading materials	No. altogether	52	124	80
	No. added by students	26 ¹	86	55
	Avg. no. added per student	0.8 ¹	1.6	1.5
Ratings	No. of ratings	181 ²	302	144
	No. of materials rated	47 (90%)	85 (69%)	51 (64%)
	Avg. no. of ratings per material	3.5	2.45	1.8
	Avg. no. of ratings per student	5.6	5.5	3.9

6. RATINGS IN USE

The possibility of rating materials was viewed positively in all three years (Table 3). Students saw ratings as highlighting which materials were worth reading and, perhaps more emphatically, which ones were not: ‘*Based on stars you could pick the best ones faster, or at least pass by the worst ones*’ (2011). Importantly, a number of students felt that 1) knowing that materials were to be rated encouraged trying to add good ones, and 2) having to rate materials made them read materials more carefully. Being able to rate materials also allowed for participating, giving and receiving feedback, and students were interested in ‘*what kinds of materials others*

Table 5: Materials views and their correlations (¹ Counting 5-3 star ratings as positive; if only 5-4 star ratings are considered positive, the correlation is $r(78) = .56, p < .01$)

		2009	2010	2011
No. of material views	Total for materials	221	674	331
	Avg. per material	4.3	5.4	4.1
	Avg. per student	6.9	12.3	8.9
Correlation between no. of views and...	...no. of ratings	$r(50) = .79, p < .01$	$r(122) = .81, p < .01$	$r(78) = .67, p < .01$
	...no. of pos. ratings	$r(50) = .76, p < .01$	$r(122) = .79, p < .01$	$r(78) = .65^1, p < .01$
	...no. of neg. ratings	$r(50) = .28, p < .05$	Not significant	Not significant
	...avg. of ratings	-	-	$r(78) = .55, p < .01$

recommended' (2010). In fact, students saw being able to participate as a value in itself: *'Of course you've got to be able to have your say'* (2010).

6.1. Selecting items for viewing

Many students felt that especially low ratings influenced their item-selecting: *'If a link had low ratings from several users, I decided not to waste my time and selected one with a better rating or without ratings'* (2011). Interestingly, however, viewing statistics do not entirely support this (Table 5). Negative ratings and the number of views a material had did not have statistically significant correlation in 2010–2011 and only a very minor one in 2009. Moreover, the 2009 correlation is positive: increased number of negative ratings correlated with increased views. For positive ratings, the correlation is highly significant in all three years.

Still, the most significant correlation, in all three years, is for the number of ratings. In fact, as rating were highly positive—over 90% in all three years if we consider three stars and above as positive in 2011 (as suggested in Cosley et al. (2003))—the high correlation between views and positive ratings may simply mean that positive ratings appear to correlate better because they were much more numerous than negative ones and that the number of ratings (volume) was decisive, not their valence.

Then again, correlations can go both ways: the correlation between the number of ratings and number of viewings can mean either that rated links—most rated positively—attracted student attention (resulting in viewings and more ratings) or that certain links attracted students irrespective of ratings (viewings resulting in ratings), for example their title interested students, perhaps matching their current information need. Student comments suggest that both happened and that there was some interplay between the two. In fact, item selecting was in many ways affected by various inter-influencing factors, for example *'The only factors that influenced me were the interestingness of the article based on the title and the number of ratings'* (2011).

In practice, valence and the number of raters formed together an important heuristic, as many students mentioned having avoided links with many

negative evaluations and selecting links with many positive ones. The importance of valence is also evidenced by a positive and statistically significant correlation between the number of star-ratings and the average of star-ratings, $r(78) = .71, p < .01$, in 2011: the higher the average of ratings, the more ratings. The likely explanation is that better rated items attracted viewings that resulted in ratings.

For some students, valence was not that central: *'That a material had stars at all made it interesting for me—I didn't pay much attention to starless ones...'* (2011). Furthermore, even negative valence could arouse interest: *'Stars were significant especially if some material had been rated as bad—in such cases I could check it out just out of curiosity'* (2011). Therefore, valence did play a role for many (often in conjunction with other factors), but its impact was not necessarily straightforward. Moreover, both valence and volume could lose their importance if other factors rendered the item uninteresting: *'But the title itself was also important. If the topic it shows isn't interesting to me, then no amount of useful votes helps'* (2010).

Some students claimed to have selected materials solely based on the title. However, even then ratings were not necessarily ignored: *'I opened links based on the title but I was very much aware of the rating'* (2011). Furthermore, social aspects of ratings were important: *'It was more interesting to read an already rated material and see if I agreed or not'* (2011). For example, one student mentioned having gone back to see how others had rated the materials he had read. In fact, participation appears to have been a goal in itself at least for some: *'...interesting to follow if others had read a material I had added or evaluated it.'* (2011)

The context also affected student behaviour, as some decided to read unrated materials on purpose: *'...I also tried to view and rate links that had not been rated yet'* (2011). In effect, some mentioned altruism as at least partial motivation: *'I tried to familiarize myself with a material for real before evaluating it and that way benefit other students, too, in addition to myself'* (2010).

In summary, students approached item selecting from multiple perspectives. While many claimed to

have used ratings mainly to identify potentially salient items or to avoid non-salient ones, there were also distinctly social undertones evident in many comments. Explaining the correlation between ratings and viewings clearly requires considering the equation from multiple perspectives.

6.2. Trust issues

Trust is important to recommender systems, as users need to be able to trust the recommendations (Swearingen and Sinha, 2001). In fact, it may have been the lack of trust on individual ratings that led some students to emphasize the number of ratings/raters as a signifier of quality: *'...if you have to rate a certain number of articles, the easiest way is to go there and click "Yes" as many times as you need to ... so ratings don't necessary mean much. Still, if an article has a large number of Yes/No ratings, it may still tell something about its quality'* (2010).

While trust issues were not serious enough to cause major distrust among the majority of students, some nevertheless questioned if everybody actually read the materials they added. Students felt that making it compulsory to write *'a short description... to make sure that the sender is familiar with the material'* (2010) when adding a material link would help make sure that students added only materials they had actually read. Similarly, some students suggested that ratings were often made dishonestly: *'...probably quite a few just clicked good reviews without actually reading the material'* (2011). Again, they felt that that commenting should be a part of a rating to make it harder to rate a material without actually reading it: *'If one had to add a textual justification ... it would've encouraged more careful reading. ...simply giving stars is perhaps too "easy"'* (2011). As a result, building trust on recommendations requires special consideration when compulsoriness is used, as the case in formal e-learning often is

Anonymity was seen as exacerbating the situation: *'Anonymity made it easy ... to add links without having to think too much if it's good or not'* (2010). Students felt that using real names would have improved the quality of added materials and the reliability of ratings. Also, students felt that anonymity reduced possibilities of social interaction—*'The full anonymity takes interactivity out of activity and, consequently, it loses part of its meaning'* (2011)—and made it impossible to judge a contribution by the person making it: *'...for example nickname X has added good comments/materials → I'll follow him more closely'* (2010). Several students suggested using nicknames so that students would have partial anonymity while still having individual presence in the system, thus improving sociality and encouraging better contributions.

6.3. Star ratings vs. binary ratings

Replacing binary ratings with star ratings clearly reduced the average number of ratings per student, largely by increasing the percentage of students who rated no items and by reducing the percentage of students who rated clearly more than the required number of items (Table 6). The difference between the means of ratings per student between 2010 and 2011 is statistically significant, $t(90) = 2.280$, $p < .01$, as is the difference between 2009 and 2011, $t(67) = 2.579$, $p < .01$, while the difference between 2009 and 2010 is not.

Table 6: Ratings by year

	2009	2010	2011
No. of ratings	181	302	144
No. of materials rated	47 (90%)	85 (69%)	51 (64%)
Avg. no. of ratings per student	5.6	5.5	3.9
No. of students with no ratings	1 (3%)	10 (18%)	8 (22%)
No. of students with less than 5 ratings	7 (22%)	18 (33%)	14 (38%)
No. of students with 5 or more ratings	25 (78%)	37 (67%)	23 (62%)
No. of students with 7 or more ratings	8 (25%)	25 (45%)	5 (14%)
No. of positive ratings	171 (94%)	272 (90%)	131 (91%)
No. of negative ratings	10	30	13
No. of dishonest ratings	73 (40%)	130 (43%)	38 (26%)

How positive the star-scale ratings were hinges on whether three-star ratings are considered positive or neutral. If only ratings of 4–5 stars are considered positive, then 63% were positive, thus marking a departure from the overwhelming positivity of the binary scale. However, if ratings of 3–5 stars are considered positive, then 91% were positive, thus mirroring the positivity of the binary scale.

On the other hand, given that 83% of the ratings were 2–4 stars, the five-star scale did successfully increase the granularity of the ratings. Even if we consider two-star ratings the lowest rating since no one-star ratings were made, ratings of 3–4 stars still represented 74% of the ratings. Extremes were more an exception than a rule, unlike in other domains (see below).

6.3.1 Increasing the complexity of the rating decision increases honesty—at a cost

One reason for opting for a more granular rating scale (five-star) was to discourage dishonest ratings. As Table 6 shows, the percentage of dishonest ratings fell significantly. The difference between the means of dishonest ratings per student between 2010 ($M = 3.13$; $SD = 3.48$) and

2011 ($M = 1.03$; $SD = 1.76$) is statistically significant, $t(90) = 3.392$, $p < .01$, as is the difference between 2009 ($M = 2.28$; $SD = 2.80$) and 2011, $t(67) = 2.262$, $p < .01$, while the difference between 2009 and 2010 is not.

The reduced number of ratings coincides with a lower correlation between ratings and number of views for star-ratings (2011) than for binary ratings (2009–2010) (Table 5). As many students saw the number of ratings as at least a partial indicator of quality, the lower number of ratings may be one of the reasons for the lower correlation.

However, the complexity of decision likely played a role here, too. With binary ratings, ratings were largely positive: in 2010, only 9 out of 124 materials (7.2%) and in 2009, 4 out of 52 (7.7%) had overall negative ratings, making ratings effectively ‘seal[s] of approval’ (Rajaraman, 2009). Conversely, the five-star scale forced students to make more complex decisions in decoding the meaning of the rating. As Figure 2 shows, the lowest rating any material had was two stars (five materials). How bad is a two-star rating—especially given that in three cases, the rating was by one student and in two by two students? Is a material rated at three stars worth checking out? The decision is probably influenced by the topic (title) and current user need. How strong recommendation is a five-star rating by one student? After all, the number of ratings was an important consideration: ‘A link with many positive evaluations was usually interesting’ (2011). Consequently, determining how good the ratings implied the material to be was more complex in 2011 than in 2009–2010.

Figure 2: Rating averages by the number of materials having the average (2011)

Student comments also indicate that students were wondering how others understood the five-star scale: ‘You never know what kind of scale and standard different people had’ (2011). What qualifies as for example four-star material may differ from one student to another. Consequently, students felt that ‘A star rating alone doesn’t tell enough about the quality of the link content’ (2011) although it can provide ‘a rough estimate’ (2011). Students felt that they needed to know the bases for the rating: ‘A star rating ... only tells roughly

how the rater saw the material. Commenting allows the rater to tell why he saw it that way’ (2011). In effect, users can better evaluate the relevance of a rating to themselves when the thinking behind it is made clear.

Thus, while increased granularity enabled students to rate materials more exactly, it simultaneously made understanding ratings more complex. In contrast, largely self-explanatory binary scale did not cause such problems in 2009–2010. While in 2009–2010 some students mentioned wishing that commenting one’s rating would have been compulsory, in 2011 the issue was mentioned by clearly more students. This implies at least two things. First, when making a binary rating in 2009 and 2010, students were asked to rate the ‘usefulness’ of the material; they were told what aspect to rate. In 2011, they were simply told to evaluate the link; the aspect to be evaluated was not stated. This underlines the importance of making clear what aspect is being rated. Second, the complexity of the decision-making process, when faced with averages of ratings on a granular scale by a number of raters, requires additional information as to why a certain rating was made to determine its relevance, especially if the volume is low and cannot therefore used as a heuristic.

Consequently, the design decision to separate rating and commenting needs to be re-evaluated, especially since coupling rating with a short comment may also encourage more careful readings: ‘If, in addition to stars, one had to add a textual evaluation, it would’ve encouraged deeper reading’ (2011). Having students state their reasons for ratings would also likely further reduce the number of dishonest ratings, as faking reasons without knowing anything about the material is difficult. On the other hand, increasing the cost of evaluation may also result in few ratings overall.

6.3.2. Five-star ratings in e-learning vs. other domains

The views of how many stars were necessary to make a material appear salient were likely influenced by student experiences in other environments. In e-commerce, ratings tend to be J-shaped (Hu et al., 2009). Also, YouTube recently switched from five-star scale to binary one because ‘the overwhelming majority of videos ... have a stellar five-star rating’—while there were some one-star ratings, two to four star ratings were very rare (Rajaraman, 2009). Rajaraman (2009) concluded that ‘the ratings system is primarily being used as a seal of approval, not as an editorial indicator...’ Most consumer-generated ratings follow YouTube’s distributional pattern: ‘rah-rah ratings’ abound, there are some low ratings and little between (Kadet, 2007).

The overall distribution of five-star ratings here did not follow the J-shaped pattern. Instead, the ratings were much more normally distributed, with emphasis on the middle values (Table 7). Only five materials had two stars (the lowest rating given) and only five had five stars as the average rating. The reason for no one-star ratings may be related to the fact that students formed a small community: *'Still, the fact that we were a tightly demarked group of users affected the ratings a lot...'* (2011), resulting in certain constraint: *'I didn't have the gall to be too critical...'* (2010).

Table 7: Distribution of ratings in 2011

	1 star	2 stars	3 stars	4 stars	5 stars
No.	0	13	42	64	25
%	0%	9%	29%	44%	17%

The reason why ratings did not follow the brag and moan model (Hu et al., 2006) may at least partially be explained by different domains (Kadet, 2007; Sparling and Sen, 2011). Love-it-or-hate-it models that have explanatory power in e-commerce are unlikely to have played a great role here: *'I usually only comment online when I really have something concrete to say or have a strong opinion on the issue. Many materials were useful but I can't say they aroused burning emotions in me...'* (2011). In addition, compulsoriness in a sense democratizes ratings, as it is not only the highly opinionated users who rate items, as for example in e-commerce (Talwar et al., 2007).

6.3.3. Characteristics of dishonesty in binary and five-star scales

On the binary scale, while ratings were overall widely positive, the dishonest ones were slightly more so. Moreover, dishonest ratings followed the previous votes slightly more faithfully: for materials with previous rating(s), 98% of the dishonest ratings followed the existing rating in 2009 and 97% in 2010, while the respective percentages for honest ratings were 92% and 90%. Thus, when making honest ratings, students tended to be slightly more critical and deviate slightly more likely from the existing evaluation.

Especially in case of honest ratings, it can be questioned whether the existing ratings affected the rater or whether the rater would have rated the item that way anyway. However, given the existing literature (for example, Cosley et al. (2003) and Lam and Riedl (2004)) on the biasing, or feel-forward, effect of prior ratings and that students reported that the existing ratings affected them—*'I evaluated materials that others had "liked" to be useful. The opinions of others affected.'* (2010)—it appears that at least some anchoring effect took place: *'I tried not to consider the ratings of others*

when rating a material but that's naturally very difficult' (2010).

Similarly, on five-star scale, for materials with previous rating(s), 70% of the dishonest ratings were the same as the existing average (i.e. within half-a-star; ratings were made with whole stars but averages displayed at the accuracy of half-a-star) while only 57% of the honest ones were. Thus, in 2011, the tendency to follow the existing rating was also stronger for dishonest raters while honest raters were readier to disagree. Also, the extremes (five and two stars in practice) represented only 21% of the dishonest votes but 38% of the honest votes. Dishonest raters appeared to have preferred middle-of-the-road ratings while honest raters were readier to take more polarizing stances.

In any case, since motivations for dishonest ratings are different in e-commerce and in e-learning—the purpose in e-commerce is to distort the aggregate rating favourably or unfavourably for an item (Lam and Riedl, 2004) while in e-learning, it is to get points without earning them honestly (Leino, 2011)—it seems unlikely that similar patterns would emerge for dishonest ratings. Following current rating trends and using middle-of-the-scale ratings (five-star) would not work for distorting aggregate ratings (shilling) but it may seem like a relatively innocuous and not easily noticeable way of getting points dishonestly.

Overall, there was a clear tendency among students making dishonest ratings to go with the flow and to avoid extremes on a granular scale and go for the positive rating on a binary one.

7. DISCUSSION: TWO MODELS

Two conflicting aspects emerge from the data, perceived trustworthiness of recommendations and the number of evaluations. The higher the cost of evaluating an item, the fewer the evaluations, but, simultaneously, the more trustworthy the evaluations were considered by the students in our study. This perception is, in fact, founded on reality, since increasing evaluation cost reduced dishonesty. Nevertheless, there needs to be a large number of evaluations (a) to achieve necessary coverage, and (b) because students use the number of ratings (along with valence) as an indicator of item quality. Regrettably, having both high-cost evaluations and a large number of evaluations appears challenging in the light of our results. This leads to two tentative models, *low-cost approach* and *high-quality approach*.

Low-cost approach emphasizes maximizing the number of evaluations. Since our results suggest that the lower the cost is, the more ratings are made, binary approach could even be ditched in favour of unary approach, for example a Like

button. The low-cost approach would encourage reading materials, as there would be numerous (volume) positive (valence) ratings—ideally many on good materials and few or none on less useful ones. After all, mostly positive binary ratings correlated clearly better with viewings than did star-rating averages. Socially, the low-cost approach might create ‘buzz’ around liked items, and thus perhaps also encourage commenting and discussing in addition to reading (Leino, 2011). However, due to high dishonesty, using such ratings for such algorithmic approaches as collaborative filtering would not work.

High-quality approach, in contrast, emphasizes improving the perceived trustworthiness and actual honesty of evaluations. Each rating would need to be accompanied by for example a comment so that users can 1) trust that the rater knows something about the item, and 2) see the reason why the rater gave the rating. The latter is important because it allows users to evaluate the relevance of the rating for themselves based on the rater (for example, whether or not similarly positioned to the item as the user) and the rating angle (for example, what aspect of the item is being rated). In e-commerce, these factors have been found significant for users in using reviews to select items (Leino and Rähkä, 2007). However, fewer evaluations are made when the rating cost is high, and so this approach would likely require more strict compulsoriness than the low-cost approach. The high-quality approach could also increase sociality around the items by generating ‘buzz’ around certain items, whether liked or contested. Moreover, the ratings could also be used for collaborative filtering or similar approaches.

Perhaps the most important benefit of the high-quality approach would be the higher reliability of recommendations—students would be directed to the best materials. In contrast, the low-cost approach would likely increase social cues in the interface and encourage viewing items without necessarily directing students to the best materials because of dishonesty.

However, neither approach is inherently better; they simply offer different approaches to using non-algorithmic recommending in e-learning. The question is what kind of activity we wish to engender and what role the ratings/recommendations play in the overall pedagogical picture.

8. CONCLUSION

Overall, students appreciated having non-algorithmic recommender features. Not only did the features make the material list more ‘*interesting*’ (2010)—‘...bring a bit of life to a lifeless list of links’ (2009)—but they were also seen as useful. Moreover, students appreciated the social and

interactional aspects that the features brought along.

The most challenging aspects were the lack of trust in the motivations of other students and the actual amount of dishonesty. Our results indicate that increasing the cost of rating increases honesty but at the cost of decreasing the number of ratings. In fact, students perceived text-based recommendations as better because (a) they are harder to fake than click-a-ratings, and (b) they let others see the reason for the rating, thereby allowing others to judge if the rating is relevant to them.

The interplay of higher honesty through increased cost of rating and lower number of contributions (and vice versa) suggested two models, *low-cost approach* and *high-quality approach*. The low-cost approach is based on lowering the cost of ratings to increase coverage and to create a ‘buzz’ around items, encouraging students to read and interact. The high-quality approach, in turn, requires ratings to be accompanied with text evaluations, thus increasing the cost of rating. Also, the rating scale employed should be granular, for example five-star rating scale, imposing a more complex decision process. The result is more reliable ratings that are also perceived as more reliable/trustworthy. However, given that the number of ratings is lower when the cost of rating is higher, in e-learning context, judicious use of compulsoriness is necessary to achieve sufficient coverage.

In the next iteration of the LSRM page, we plan to use evaluations where a short comment is a compulsory part of a five-star rating to see how making evaluating more costly affects honesty, number of ratings and perceived trustworthiness of ratings, and to replace anonymity with nicknames to see how this affects perceived social presence and perceived quality of materials and evaluations.

In e-learning, an important task for recommender features is to encourage desired behaviour in addition to helping students find and select items. In our case, we wanted students to read more high-quality materials, selected and evaluated by a community of peers. Inducing desired behaviour requires selecting the right set of recommender features, and selecting the right set of features requires understanding the underlying dynamics of the ecosystem. This exploratory study reveals some of these dynamics and provides practitioners with actionable information.

9. ACKNOWLEDGMENTS

This work was partly supported by the Academy of Finland grant 129335. The author wishes to thank Professor Kari-Jouko Rähkä for his comments.

10. REFERENCES

- Cena, F., Likavec, S., Lombardi, I. and Picardi, C. (2011) Synthesis of collective tag-based opinions in the socialweb. *LNCS*, 6934. 286-298.
- Cosley, D., Lam, S.K., Albert, I., Konstan, J.A. and Riedl, J. (2003) Is seeing believing?: how recommender system interfaces affect users' opinions. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Ft. Lauderdale, FL, 5-10 April 2003. ACM: New York, NY. 585-592.
- Drachsler, H., Hummel, H. G.K. and Koper, R. (2009) Identifying the goal, user model and conditions of recommender systems for formal and informal learning. *Journal of Digital Information*, 10 (2). 4-24.
- Gena, C., Brogi, R., Cena, F. and Venero, F. (2011) The impact of rating scales on user's rating behaviour. *LNCS*, 6787. 123-134.
- Ghauth, K.I. and Abdullah, N.A. (2010) Learning materials recommendation using good learners ratings and content-based filtering. *Educational Technology Research and Development*, 58 (6). 711-727.
- Hage, H. and Aïmeur, E. (2008) Harnessing learner's collective intelligence: A Web2.0 approach to e-learning. *LNCS*, 5091. 438-447.
- Hennig-Thurau, T., Gwinner, K.P., Walsh, G. and Gremler, D.D. (2004) Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, 18 (1). 38-52.
- Herlocker, J.L., Konstan, J.A., Terveen, L.G. and Riedl, J.T. (2004) Evaluating collaborative filtering recommender systems. *TOIS*, 22 (1). 5-53.
- Hu, N., Pavlou, P. A. and Zhang, J. (2006) Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication. In: *Proceedings of the 7th ACM conference on Electronic commerce*, Ann Arbor, MI, 11-15 June 2006. ACM: New York, NY. 324-330.
- Hu, N., Zhang, J. and Pavlou, P.A. (2009) Overcoming the J-shaped distribution of product reviews. *Commun. ACM*, 52 (10). 144-147.
- Kadet, A. (2007) *Rah-rah ratings online*. Available from <http://www.smartmoney.com/spend/rip-offs/rah-rah-ratings-online-20837/> (8 Feb. 2013).
- Konstan, J., & Riedl, J. (2012) Recommender Systems: From Algorithms to User Experience. *UMUAI*, 22. 101-123.
- Lam, S.K. and Riedl, J. (2004). Shilling recommender systems for fun and profit. In: *Proceedings of the 13th international conference on World Wide Web*, New York, NY, 17-22 May 2004. ACM: New York, NY. 393-402.
- Leino, J. (2011) Case study: material additions, ratings, and comments in a course setting. In: Santos, O.C. and Boticario, J.G. (Eds.), *Educational Recommender Systems and Technologies: Practices and Challenges*, IGI Global: Hershey, PA. 258-280.
- Leino, J. (2012) Case study: Recommending course reading materials in a small virtual learning community. *IJWBC*, 8 (3), 285-301.
- Leino, J. and Räihä, K.-J. (2007) Case Amazon: ratings and reviews as part of recommendations. In: *Proceedings of the 2007 ACM conference on Recommender systems*, Minneapolis, MN, 19-20 Oct. 2007. ACM: New York, NY. 137-140.
- Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H.G.K. and Koper, R. (2011) Recommender Systems in Technology Enhanced Learning. In: Kantor, P.B., Ricci, F., Rokach, L. and Shapira, B. (Eds.), *Recommender Systems Handbook*, Springer: Berlin, Heidelberg. 387-415.
- Rajaraman, S. (2009) *Five stars dominate Ratings*. Available from <http://youtube-global.blogspot.com/2009/09/five-stars-dominate-ratings.html> (8 Feb. 2013).
- Schafer, J.B., Konstan, J. and Riedl, J. (1999) Recommender systems in e-commerce. In: *Proceedings of the 1st ACM conference on Electronic commerce*, Denver, CO, 3-5 Nov. 1999. ACM: New York, NY. 158-166.
- Sparling, E.I. and Sen, S. (2011) Rating: how difficult is it? In: *Proceedings of the 5th ACM conference on Recommender systems*, Chicago, IL, 23-27 Oct. 2011. ACM: New York, NY. 149-156.
- Swearingen, K. and Sinha, R. (2001) Beyond algorithms: an hci perspective on recommender systems. In: *Proceedings of the ACM SIGIR Workshop on Recommender Systems*.
- Talwar, A., Jurca, R. and Faltings, B. (2007) Understanding user behavior in online feedback reporting. In: *Proceedings of the 8th ACM conference on Electronic commerce*, San Diego, CA, 4-11 June 2007. ACM: New York, NY. 134-142.
- Tang, T. and McCalla, G. (2004) Beyond learners' interest: personalized paper recommendation based on their pedagogical features for an e-learning system. *LNCS*, 3157. 301-310.